---
title: "NYPD_Shooting_Project"
author: "Ryan Scott"
date: "2025-05-31"
output:
  pdf_document: default
  html_document: default

R Markdown

NYPD Shooting Incident Project

Statement of Interest

The following project is an Exploratory Data Analysis of NYPD Shooting Incident Data from 2006 to 2024. The questions we want to answer are:

What year had the highest number of shooting incidents? Which Boroughs had the highest number of shooting incidents?

Background

We were able to access the files for this project at https://catalog.data.gov/dataset and find the dataset titled NYPD Shooting Incident Data (Historic).

To begin, I install the appropriate R packages.

```
{r} install.packages("tidyverse")

library(tidyverse)
```

I then assign the NYPD csv file to a variable in Rstudio.

```
{r}
NYPD <- read_csv("~/Desktop/Documents/Data Analysis Tools/Projects/RStudio Projects/NYPD/NYPD_Shooting_Incident_Data__Historic_.csv")
```

I want to get an idea of the columns I will be working with, so I use the head function and set the number of rows to 5.

```
{r}
head(NYPD, n = 5)
```

I use the summary function to identify the types of data in each column.

```
{r}
summary(NYPD)
```

After reviewing the columns and the data types of the columns, I change Occurrence Date to date format from character and remove the Latitude, Longitude, and Coordinate columns from the table

```
{r}
NYPD <- NYPD %>%
   mutate(OCCUR_Date = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))
```

```
{r}
NYPD <- NYPD %>%
    select(-Latitude, -Longitude, -Lon_Lat, -Y_COORD_CD, -X_COORD_CD)
```

I want to see the date range of the data, so I check the minimum and maximum dates of the Occurrence Date column.

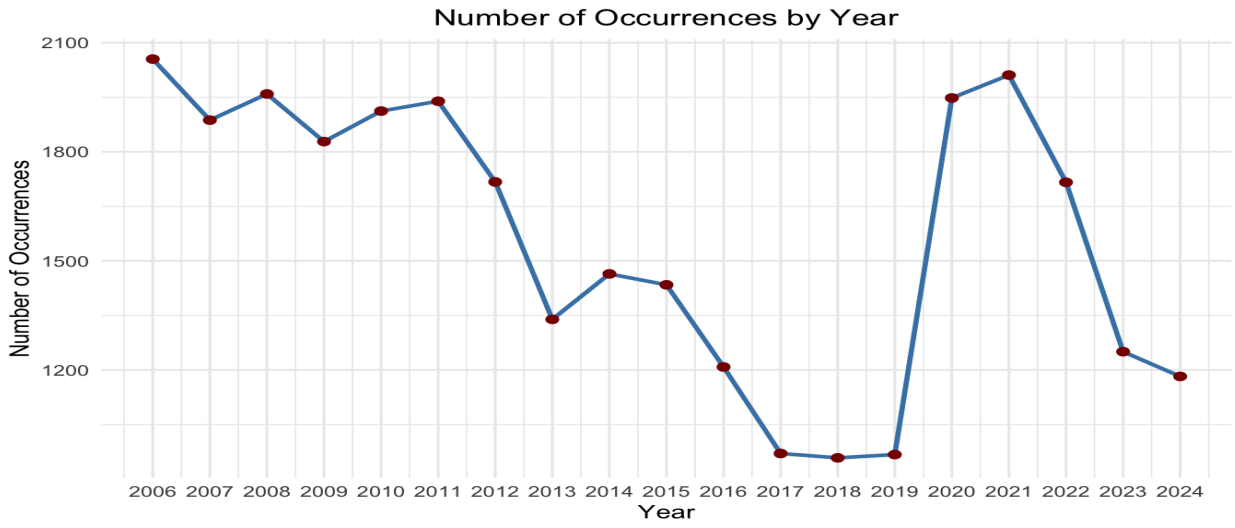
```
{r}
date_summary <- NYPD %>%
    summarise(
        max_occur_date = max(OCCUR_Date, na.rm = TRUE),
        min_occur_date = min(OCCUR_Date, na.rm = TRUE)
    )

print(date_summary)
```

From this code, I was able to determine that the date range is from 01-01-2006 to 12-31-2024.

The following visualization shows the number of occurrences by Year in ascending order.

```
{r occurences_by_year}
ggplot(occurrences_by_year, aes(x = Occur_Year, y = Count)) +
    geom_line(color = "steelblue", size = 1) +
    geom_point(color = "darkred", size = 2) +
    labs(
        title = "Number of Occurrences by Year",
        x = "Year",
        y = "Number of Occurrences"
    ) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_x_continuous(breaks = unique(occurrences_by_year$Occur_Year))
```

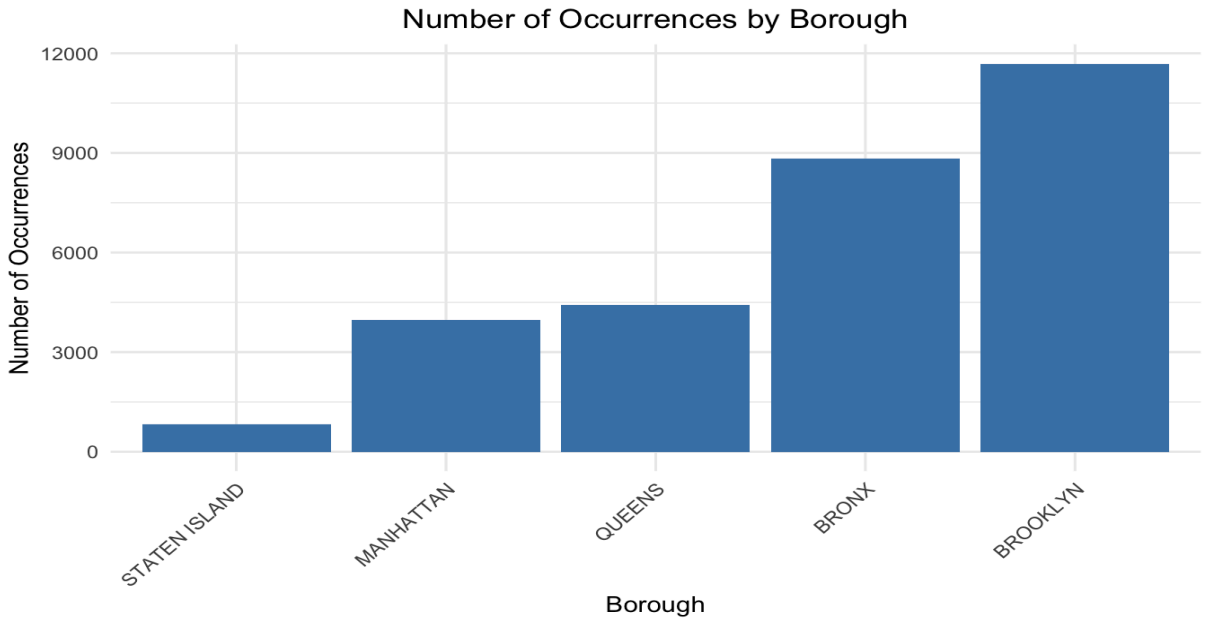


From the previous visualization, I am able to determine that with this date range, 2006 had the highest number of occurrences.I am also able to determine that number of occurrences fell from 2006 to 2019 before rising sharply in 2020.

The following visualization shows the number of occurrences by New York City Borough.

```
{r borough_counts}
ggplot(borough_counts, aes(x = BORO, y = n)) +
    geom_col(fill = "steelblue") +
    labs(
        title = "Number of Occurrences by Borough",
        x = "Borough",
```

```
    y = "Number of Occurrences"
) +
theme_minimal() +
theme(
    plot.title = element_text(hjust = 0.5), # Center the title
    axis.text.x = element_text(angle = 45, hjust = 1)
)
```



After reviewing the previous visualization, I can determine that within the date range of 2006 to 2024, Brooklyn had the highest number of occurrences.

Conclusion

Upon conclusion, we were able to determine that 2006 had the highest number of NYPD Shooting Incidents and that Brooklyn had the highest number of shooting incidents.

Bias

I could not identify any source of bias within this project. I kept the questions simple and open ended to elimate any possible bias. As far as the data goes, the columns related to facts about the incidents, such as time, location, and date, so bias would not be present.